Fine Scale Genomic Signals of Admixture and Alien Introgression among Asian Rice Landraces

João D. Santos[1,2], Dmytro Chebotarov[3], Kenneth L. McNally[3], Jérôme Bartholomé[1,2,3], Gaëtan Droc[1,2], Claire Billot[1,2], and Jean Christophe Glaszmann[1,2,*]

[1]UMR AGAP, CIRAD, Montpellier, France

[2]UMR AGAP, Université de Montpellier, France

[3]International Rice Research Institute (IRRI), Los Baños, Philippines

*Corresponding author: E-mail: glaszmann@cirad.fr.

Accepted: April 11, 2019

Abstract

Modern rice cultivars are adapted to a range of environmental conditions and human preferences. At the root of this diversity is a marked genetic structure, owing to multiple foundation events. Admixture and recurrent introgression from wild sources have played upon this base to produce the myriad adaptations existing today. Genome-wide studies bring support to this idea, but understanding the history and nature of particular genetic adaptations requires the identification of specific patterns of genetic exchange. In this study, we explore the patterns of haplotype similarity along the genomes of a subset of rice cultivars available in the 3,000 Rice Genomes data set. We begin by establishing a custom method of classification based on a combination of dimensionality reduction and kernel density estimation. Through simulations, the behavior of this classifier is studied under scenarios of varying genetic divergence, admixture, and alien introgression. Finally, the method is applied to local haplotypes along the genome of a Core set of Asian Landraces. Taking the Japonica, Indica, and cAus groups as references, we find evidence of reciprocal introgressions covering 2.6% of reference genomes on average. Structured signals of introgression among reference accessions are discussed. We extend the analysis to elucidate the genetic structure of the group circum-Basmati: we delimit regions of Japonica, cAus, and Indica origin, as well as regions outlier to these groups (13% on average). Finally, the approach used highlights regions of partial to complete loss of structure that can be attributed to selective pressures during domestication.

Key words: 3,000 Rice Genomes, *Oryza sativa*, SNPs, local haplotypes, population structure, kernel density estimation.

Introduction

High throughput genotyping technologies provide at low cost an increase of marker density and sampling sizes. They have changed considerably the way we study the genetic structure of populations. The benefits of discriminating between contrasting local signals across genomic locations were quickly appreciated in association studies, where the correction for population structure evolved to become locus specific (Diao and Chen 2012; Park et al. 2015). In population genetics, this increase in precision brought with it a realization of the pervasiveness of admixture in natural populations. Instead of providing coherent indicators of population membership, the genomes of admixed individuals consist of fragmented contributions from differentiated populations. The relation of traditional descriptors of genetic variation, such as F_{st} and principal component analysis (PCA) (Slatkin 1991; Novembre and Stephens 2008), to mean coalescent times across markers, limits their utility in these cases. By focusing on small enough regions of the genome, researchers now have the possibility of analyzing these contrasting signals separately. As a result, the effects of migration are increasingly captured as a by-product of the description of genetic structure in the form of local ancestry assignments of hybrid individuals (Gravel 2012; Henn et al. 2012). The literature on this subject has grown rapidly in recent years, with multiple approaches proposed (Gravel 2012; Geza et al. 2018). This development is appreciated particularly in the study of species of agronomic interest, where a fine genetic resolution is needed for searching and combining favorable genes.

Since its domestication in Asia, rice has evolved a broad range of diverse cultivars adapted to multiple environments and regional food preferences (Khush 1997). The oldest

systematic descriptions of rice diversity come from China, where two major types were already recognized two millennia ago, namely Hsien—or Xian, early-ripening rice—and Keng—or Geng, late-ripening soft-cooking grained rice (Ho 1956; Oka 1988). Early classifications confirmed the prominence of this scheme beyond China, and the two types were renamed Indica and Japonica, respectively (Kato 1928; Matsuo 1952). A third group, named Javanica, was temporarily used (Morinaga 1954) but was eventually recognized as a tropical form of the Japonica (Glaszmann and Arraudeau 1986). A study with broader and more systematic ecologic and geographic coverage using biochemical markers led to the recognition of additional clusters, such as that containing the Aus varieties from the Indian subcontinent and another including the famous Basmati varieties (Glaszmann 1987). Subsequent work with more markers confirmed these groups and added subgroups within the Japonica (Garris et al. 2005). The binary Indica–Japonica scheme remained a central feature in the architecture of rice classification until recently, when a study by Huang et al. (2012) placed Japonica as the original domesticate and attributed the Indica group to the introgression of Japonica domestication alleles into other genetic backgrounds. This scheme identified the Aus group as derived from Indica and posited Basmati-similar varieties to be a direct derivation of the Japonica. However, a recent reinterpretation of the same data concluded on three sources of domestication, distributed across the Indica, Japonica, and Aus, and the Basmati-like cluster to be a hybrid group between Aus and Japonica (Civáň et al. 2015). The same data were also used to reveal massive gene flow between cultivated and wild rice, posited to have led to the "feralization" of most wild rice (Wang et al. 2005, 2017). The most recent large-scale analysis (Wang et al. 2018), based on the sequences of over 3,000 rice cultivars, revealed a structure featuring Indica and Japonica, both with multiple subgroups with high geographical coherence, and concluded that rice resulted from multiple domestications, confirming the reality of circum-Aus (or cAus, the group that includes the Aus ecotype) and circum-Basmati (or cBasmati, the group that includes the Basmati types) as varietal groups. Today it is clear that the diversity of rice rests on more than one foundation event with profuse genetic exchanges among cultivated forms and between them and their wild relatives. It is likely that as data sets grow in size additional sources of genetic diversity will be uncovered.

In this context, the genome of rice cultivars can be thought of as a mosaic of segments with different origins and histories related to both primitive domesticates and occasional wild contributors. This interwoven state of different branches of variation poses a serious challenge to the study of the history of modern rice cultivars and is at the root of the controversy among early studies based on global structure analyses. An intuitive and informative description of this variation is thus crucial for our understanding of rice history and the exploitation of this data for agronomic purposes. We make use of the 3,000 Rice Genomes data set (3K RGP 2014; Wang et al. 2018) to study the patterns of exchange and overlap between globally structuring elements of domesticated rice diversity (*Oryza sativa* L.). This exploration was guided by four requirements: 1) the adoption of an initial reference scheme that best reflects current views on the genetic foundation of modern rice cultivars, 2) the identification of the most likely origin of local haplotypes along the genome of a large sample of cultivated rice, 3) the identification of exchanges of genetic material among major varietal groups, and 4) the identification of contributions to cultivated varieties from branches of variation alien to those groups. Although the study of rice genetic structure is facilitated by the near-complete homozygosity of rice genomes, these requirements place the goal of our analysis outside the scope of existing software of local ancestry inference. We developed an approach tailored to this data set and to our objectives.

Our presentation is structured as follows. We first describe a method that allows us to place local haplotypes relative to the distributions of reference populations reliably (throughout the article, the terms *global* and *local* are used in a genomic sense, not a geographic sense). Through simulations, the response of this classifier to varying degrees of genetic structure is explored. In order to highlight the need for a tailored approach, the performance under the same scenarios of the most recent software developed for similar descriptions (Dias-Alves et al. 2018) is also studied. Next is described the application of the resulting custom classification method to the 3,000 rice genome data set. Because the description of genomic variation proposed is dependent on the choice of reference populations, a mixture of genetic analysis and prior knowledge was employed in order to select, from the over 3,000 rice cultivars in the 3K RG, a subset of landrace varieties. This subset includes three Core Reference Groups (CRGs; Indica, Japonica, and cAus) and samples to be more deeply studied. Finally, the results of the application of the proposed protocol to the curated rice data set are discussed in light of their implications to rice history and breeding.

Materials and Methods

Materials

The 3K Rice Genomes is the largest plant genomics data set available today. It includes over 3000 resequenced rice genomes aligned to the reference genome of *O. sativa* ssp. *japonica* cv. Nipponbare genome. This data set contains over 29 million biallelic markers. The unfiltered data set was downloaded from the SNP-Seek database during November 2016 (http://snp-seek.irri.org/; Mansueto et al. 2017).

We removed loci with over 0.1% missing data and over 5% heterozygous sites. The resulting data set includes 10,459,872 single nucleotide polymorphisms (SNPs). Mean distance among SNPs is of 32 bp, with a standard deviation

of 514 bp. We avoided minimum allele frequency and Linkage Disequilibrium filtering. First, this would eliminate rare variants, including possible exogenous contributions to infrequent but important varietal types such as the cBasmati, a group that comprises <2% of the total data set. Second, it would remove SNPs that contribute to the local regions that our analyses are geared to discover.

We performed a global structure analysis using sNMF v. 1.2 (Frichot et al. 2014, supplementary fig. S1, Supplementary Material online). At $K = 4$, we applied a threshold of 0.8 to sNMF admixture proportions in order to draw the signaled reference groups with rather stringent criteria. We identified four major groups of rice diversity, corresponding to the Japonica (GJ), Indica (Xien-Indica, XI), cAus, and cBasmati of Wang et al. (2018).

We then selected accessions to form a sample of traditional materials representative of initial crop diversity in Asia, that is, after domestication but before global dispersion and modern breeding. For each Asian country the selection included a representation of the groups observed there (groups as in Wang et al. [2018]) taking in priority landraces and excluding improved materials, except a few well-known varieties such as Nipponbare or IR64. When it was possible the quality of the sequence, evaluated with the mapping rate, was used as an additional criterion. The final material consisted of 948 accessions, of which 654 were labeled "landraces" and eight were labeled "improved," the rest being unlabeled. It included 395 Indica, 320 Japonica, 65 cAus, and 168 admixed accessions, among which 62 belonged to cBasmati (supplementary table S1, Supplementary Material online). In line with the three-pillar view of rice domestication (Civáň et al. 2015), we hereafter refer to these groups of Indica, Japonica, and cAus materials as CRGs.

Development of a Method of Assignment

The characterization of local haplotype variation relative to reference groups requires an accurate description of local genetic distributions. In view of using PCA for analyzing local genomic variation in a system as complex as rice, we began by exploring the relationship between the genetic correlation of population samples and the respective PCA feature space distances. Population samples were generated using the Beta distribution. Data sets were simulated using arbitrary population numbers, random sampling and arbitrary genetic distances between populations (supplementary Methods: Simulations, Supplementary Material online). Haplotype diversity was calculated for each population K at each locus as $H_S = 1 - (p^2 + q^2)$, with p and q the precomputed allele frequencies at that locus. Between two populations, expected heterozygosity H_T was estimated at each locus as $2\bar{p}\bar{q}$ and F_{st} values were calculated as $\frac{H_T - H_S}{H_T}$. Average F_{st} was calculated across loci between frequency vectors of simulated populations. Pairwise population Euclidian distances were calculated between the centroids of the PCA projections of samples generated for each population. Pairwise population F_{st} values and centroid distances were pooled by sample length (supplementary Methods: PCA and Genetic Distributions, Supplementary Material online). We found that for a sufficient number of principal components (PCs) and up to eight populations the correlation between genetic and feature space distances remains high across vector lengths (Pearson's $r >= 0.98$, supplementary fig. S2, Supplementary Material online). Given this relationship, we chose to estimate the probability density function of reference haplotypes in feature space. We resorted to the kernel density estimate (KDE) of reference samples in feature space.

Local Classification: Kernel Density Estimation in PCA Feature Space

Kernel density estimation, a commonly used tool in classification algorithms involving feature reduction (Silverman 1998; Scott 2015), was used to classify local haplotypes into classes corresponding to the CRGs. KDEs were extracted locally for each reference group. KDE was run on the first five dimensions of the projections of each CRG following PCA transformation of the Core data set. The function *KernelDensity* of the python package *sklearn.neighbors* (Pedregosa et al. 2011) was used with Gaussian kernels. For bandwidth estimation, the function *GridSearchCV* of the python package sklearn (Pedregosa et al. 2011) was used. The log-likelihood of each observation was estimated. To ensure the comparability of the scores derived from different KDEs—and the applicability of a single outlier threshold, reference-specific Z scores were calculated in log-space using the mean and standard deviation of those accessions used for KDE only. Their lower-tail P values were derived from this reference-specific distribution under the assumption of normality.

This method presents several advantages: 1) density measures combine the proximity and number of labeled neighbors, 2) by choosing Gaussian kernels we impose only basic assumptions on the nature of genetic differentiation, that of a cluster specific binomial probability of mutation at each site, and 3) by estimating kernel bandwidth from the distribution of pairwise genetic differences, we are supplying the above mentioned assumption with a data driven proxy of genetic differentiation. Finally, the log-likelihoods of Gaussian KDEs can be normalized, allowing for comparison and outlier identification (Aggarwal 2016).

In order to validate the application of a KDE-based classifier to genetic data in the framework of the 3K Rice Genomes data set, we first tested the capacity of this tool to discriminate reference groups in feature space. Under stationary conditions, the KDE of population samples in feature space was used to derive an estimate of the degree to which different populations can be distinguished, distribution overlap (supplementary Methods: PCA and Genetic Distributions,

Supplementary Material online), and the relation of this measure to genetic correlation was analyzed. We then explored the application of the same concept and tool to derive an association-informative classification of individual samples. Lower-tail *P* values were derived from reference population KDEs and samples assigned according to the maximum score. The relation of classification accuracy and the overlap measure was studied in scenarios of symmetric and asymmetric overlap of unimodal and multimodal distributions independently of phylogeny (supplementary Methods: Genetic Structure and Classification, Supplementary Material online). Results were compared with those published for the WINPOP method visually, and to the output of the software *Loter* (Dias-Alves et al. 2018) on simulated data. Then, by establishing a threshold on the comparison of local KDE-derived *P* values, we studied the use of intermediate classes to act as buffers for the drop in classification accuracy at lower genetic distances. For the identification of alien material, we tested the relationship between genetic distance and outlier classification using a lower-tail *P* value threshold (supplementary Methods: Outlier Identification, Supplementary Material online).

Having characterized the behavior of KDEs in a neutral scenario, we set out to test its efficiency under the particular conditions of rice data. To this end, we required a characterization of population samples in data sets of local genomic data. We calculated allele frequencies on local clusters, estimated using the unsupervised clustering algorithm mean shift (Comaniciu and Meer 2002) at 1,000 random windows of length 150. Allele frequencies were calculated at clusters of over 35 individuals. We then calculated pairwise F_{st} values between allele frequency vectors. Finally, we tested the accuracy of KDE under maximum likelihood, and the proportion of haplotypes classified as intermediate and outlier when applying thresholds using pairs of vectors along a range of genetic distances (see supplementary Methods, Supplementary Material online).

Application to Rice Core Data Set

PCA was applied to the haplotypes of the 948 Asian traditional landraces at 137,691 windows of 150 SNPs along the genome (mean physical size: 5,295.3 bp, sd = 9,995.6), the same length as used during simulations. Windows overlapped over half their own size in terms of SNP number. For each data set, the kernel density of each CRG (Japonica, Indica, and cAus) was estimated in feature space and the respective log-likelihoods extracted for each of the 948 accessions. In each case, log-likelihoods were normalized by those of CRG accessions and their respective lower-tail *P* values extracted assuming normality.

For classification, we resorted to a lower *P* value threshold for outlier assignment, and to a *P* value ratio to distinguish between pure and intermediate assignments. The lower threshold was set to 0.0001, locally assigning to the outlier class any observation whose *P* value under the three KDEs fell below this value. For the classification into pure and intermediate classes of nonoutlier observations, we studied the impact of a range of thresholds on *P* value ratios on the final output (supplementary fig. S3, Supplementary Material online). For the statistics and ideograms reported here, this threshold was set to 4, at which classification proportions are observed to reach a plateau (see Results). Under a scenario of three reference groups, this practice results in four intermediate classes, one for each pairwise comparison and one for three-way uncertainty. No smoothing across *P* values at the individual level was attempted.

Following local classification of all windows, for ideogram construction and physical summary statistics, windows were compressed by individual: in order of increasing first SNP, windows of the same classification were merged into single blocks. A new block was created with every change in class. Merged blocks range from the first SNP of the first window merged to the first SNP of the next block minus one.

We analyzed distribution of *P* value overlap between the three references across all windows covering genic regions. Genic regions were first extracted from the MSU rice genome annotation database v. 7 (Kawahara et al. 2013). Overlap was measured at each window by summing the proportions of minimum to maximum CRG-specific *P* values across pairwise combinations for each individual. The median of this overlap was taken across CRG accessions. Each gene was indexed to every window overlapping with its IRGSP 1.0 coordinates. Overlap values for gene specific windows were averaged. MeanShift clustering was applied to the resulting vector of gene overlap scores. Bandwidth was estimated from that vector and outliers were discarded.

We focused on regions surrounding some genes of interest in the study of the domestication of rice —*Bh4*, *qSH1*, *OsC1*, *Rc*, and *Sh4*—whose diversity has already been described earlier, and which cover a range of distribution overlap scenarios, to serve as an illustration of the mode of classification used. For those foci, local diversity was summarized using Neighbor-Joining as implemented in DARwin software v.6 (Perrier and Jacquemout-Collet 2006), onto which the output of local KDE-based classification was projected.

Results

Local Analysis and Classification

In preparation for the automated analysis of local genomic windows, we began by exploring the relationship between Euclidian distances in the feature space of PCA and the genetic fixation index, F_{st}, under the varying constraints of population number, marker number and sampling bias (McVean 2009). Preliminary analyses show that using five PCs the transformation of genetic distances between up to nine independent populations into PCA feature space Euclidian distances is

robust and independent of sampling bias (supplementary fig. S2, Supplementary Material online). This result motivated the use of the probability density function of reference samples in feature space for assignment. The kernel density estimation of population samples in feature space was used to derive reference-specific lower-tail *P* values. Because KDE will adjust to structured (multimodal) distributions, a measure of the degree to which two populations can be distinguished (distribution overlap) in feature space was derived and related to F_{st} and to the accuracy of the maximum-likelihood classification of genetic samples into reference groups (supplementary fig. S4, Supplementary Material online). Between two populations, the assignment of haplotypes using their maximum *P* values was found to produce a decreasing error rate relative to F_{st} and to be positively related to distribution overlap (supplementary fig. S4, Supplementary Material online). This effect was reproduced when applying the software *Loter* to the same data (supplementary fig. S5, Supplementary Material online). Finally, the distribution overlap measure was used to extend this observation to scenarios of structured, asymmetric overlap (supplementary fig. S6, Supplementary Material online). In this context, we explored the behavior, relative to F_{st}, of the classification of samples based on a threshold on the comparison of pairwise *P* values. Under thresholds of 2.0 through 6.0 and based on the comparison of normalized likelihood estimates, both the rate of miss-assignment and that of classification into an intermediate class are proportional to the degree of overlap between reference distributions (supplementary fig. S7, Supplementary Material online). At a threshold of 4.0, intermediate classification reaches a proportion of 0.95 when relative distribution overlap rises to 0.8 (supplementary fig. S7*A–C*, Supplementary Material online). As a consequence, the proportion of otherwise miss-classified samples falling into this class reaches 1 at an overlap of 0.8 (supplementary fig. S7*D–F*, Supplementary Material online). This entails a loss of information when genetic distance is low (supplementary fig. S7*G–I*, Supplementary Material online) but ensures a drop in miss-classification between reference classes (supplementary fig. S7*J–L*, Supplementary Material online).

The application of this custom method of classification into pure and intermediate classes significantly reduces the rate of miss-assignment and can provide information on the distribution of local patterns of genetic structure (supplementary figs. S8 and S9, Supplementary Material online).

For the identification of outlier material, a lower threshold of 0.0001 on maximum local *P* values derived from reference KDEs was used to explore the relation between outlier classification and genetic distance. In the case of pure reference populations this approach was found to consistently assign haplotypes from sources distant by over 0.03 F_{st} to this class (supplementary fig. S10, Supplementary Material online). However, if material from a foreign source is represented in reference samples by over a 3% margin, our results show that it will be classified into that group and not as outlier (supplementary fig. S11, Supplementary Material online). Despite these limitations, the allowance for the local identification of outliers provides an improvement over other reference based methods that lack this option. The application of *Loter* to the simulated data set of outlier introgressions among admixed samples reveals how without the allowance for a lower threshold, outlier material is constitutively assigned to one of the reference groups (supplementary fig. S12, Supplementary Material online).

Using locally estimated allele frequencies, we find the impact of correlation on accuracy and on intermediate classification proportions to be more stringent. Accuracy under maximum likelihood and intermediate and outlier classification proportions under the use of thresholds rise over 80% below the threshold of 0.02 F_{st} (supplementary fig. S13, Supplementary Material online). This increase in accuracy is accompanied by a greater proportion of variance captured within the first five PCs, which more closely resembles that obtained on real data (supplementary fig. S14, Supplementary Material online). Across real data sets and data sets generated from observed frequency vectors, 89.4 and 84.3% of total variance is retained within the first five PCs on average (7.6 and 15.9% sd, respectively). For comparison, the proportion of total variance explained by the first five PCs averages 24% (sd = 7.6%, supplementary fig. S15, Supplementary Material online) when using data sets of samples generated using the Beta distribution.

On this basis, the custom classifier provides a description of the strength of local associations. Concerning its application to local haplotype variation, to the point that recombination is rendered negligible at this level, such that genetic correlation is indicative of identity by descent, these associations should represent the shared presence along true evolutionary branches. However, because this approach does not rely on the calculation of the ancestral nature of reference groups but solely on an analysis of their local distributions, the origins of ancient exchanges of genetic material will only be recognized if they result in significant distribution overlap once having reached fixation across subsets of separate reference populations. Concomitantly, introgressions of wild material today prevalent among references will be classified accordingly. In the context of rice, this limitation is accepted given the difficulties in resolving the polyphyletic origins of modern cultivars and the patchwork nature of their genomes.

Distribution of Differentiation along the *O. sativa* Genome

Overall, sufficient structure exists along the *O. sativa* genome to discriminate between reference groups. On average, 46.2% of the genomes of Indica accessions are classed into a single pure group or another, as are 57.2% of the *c*Aus genomes and 56.2% of the Japonica genomes (table 1). The proportion of each genome assigned to intermediate classes

Table 1
Mean Percentage of Genome Assigned by Class, Using Local, KDE-Based Classification and Core Reference Groups

	Global Classification				
Local, KDE-Based Classification	**Indica (%)**	**cAus (%)**	**Japonica (%)**	**cBasmati (%)**	**Other Admix (%)**
Indica	**42.9 (34.9–49.8)**	2.4 (0.9–5.5)	1.3 (0.1–6.9)	6.1 (4.2–19.3)	17.1 (1.7–36.6)
cAus	2.1 (0.2–11.1)	**53.7 (43.9–60.7)**	0.5 (0.1–2.4)	11.7 (6.3–20.7)	13.4 (0.6–45.3)
Japonica	1.2 (0.2–7.1)	1.1 (0.2–6.7)	**54.5 (40.5–59.2)**	28.0 (15.5–34.2)	14.9 (0.3–51.6)
Jap-Ind	11.6 (8.6–13.9)	1.3 (0.3–3.0)	14 (12.0–16.0)	10.2 (8.2–11.2)	10.0 (2.6–16.1)
Ind-*c*Aus	23.0 (19.6–25.9)	19.7 (11.0–23.9)	0.8 (0.1–3.4)	6.7 (4.9–13.1)	15.1 (1.6–24.6)
Jap-*c*Aus	0.9 (0.6–2.4)	6.3 (3.8 - 9.0)	9.4 (5.5–11.1)	7.7 (5.1–9.1)	5.0 (1.0–22.9)
Jap-Ind-*c*Aus	17.6 (15.6–21.7)	15.1 (12.8–18.0)	18.6 (17.5–21.1)	17.0 (13.7–18.9)	18.3 (14.2–22.9)
Outlier	0.7 (0.0–7.1)	0.5 (0.1–1.7)	1.0 (14.0)	12.7 (9.0–25.5)	6.2 (0.4–40.1)

Bold values indicate congruent global and local classifications.
NOTE.—To estimate physical region assignment by class, SNPs were assigned as described in Materials and Methods for summary statistics, and length of local blocks was estimated as range between SNPs of different assignment. Length of local blocks was summed by class for every accession (min and max values in parentheses).

varies across reference groups. Globally, genomic coverage of intermediate cAus–Indica classifications (average 23% and 19.7% among Indica and cAus, respectively, 0.8% among the Japonica) dominates over Japonica–cAus (average 9.4% and 6.3% among the Japonica and cAus, respectively, 0.9% among the Indica) and Japonica–Indica (average 11.6% and 13.9% among Indica and Japonica, respectively, 1.2% among the cAus). Three-way intermediate classifications cover on average 17.6%, 15.1%, and 18.6% of Indica, cAus, and Japonica genomes, respectively.

Among pure classifications, a number of these contradict the global classes of the accessions that carry them. An average of 1.2% and 2.1% of the genome of Indica accessions is classified as Japonica and cAus, respectively, whereas 1.1% of cAus genomes are classified as Japonica and 2.4% as Indica and an average of 1.3% and 0.5% of Japonica genomes are classified as Indica and cAus, respectively (table 1).

Assignment proportions across admixed accessions vary widely, with pure Indica, cAus, and Japonica classification covering as much as 36.6%, 45.3%, and 51.6% of some of these genomes, and as little as 1.7%, 0.6%, and 0.3% for others, respectively (table 1). The scenario is different when considering cBasmati accessions alone, where assignment patterns appear much more conserved. Among accessions of this group, pure Japonica assignments cover an average of 27.9% of the available genome, cAus assignments 11.7%, and Indica assignments 6.1%. Outlier assignments are also particularly prominent and localized, comprising an average of 12.7% of the genomes of these accessions.

Local classifications can be displayed, ordered by genomic position and individual, in the form of chromosome specific ideograms (fig. 1). Accessions are placed vertically in the following order: Indica, cAus, Japonica, and Admixed, with cBasmati first, followed by other admixed. The ordering of window-based classifications by genomic position reveals the physical association between assignments to the same class. Figure 1 displays several cases of contradictory combinations of putative local genomic origin versus CRG classification, which are often physically clustered and shared among several accessions of the same group.

We analyzed the distribution of locally versus globally discordant assignments across genome-wide classifications (fig. 2). The physical sizes of windows assigned to the same class were summed by accession and the distribution of this variable was grouped by class across the global groups and subgroups described by Wang et al. (2018). Among Japonica subgroups, we find the tropical Japonica to present the largest mean total coverage of assignment to both Indica and cAus classes, with the temperate Japonica at the bottom of that distribution (fig. 2). Visual analysis of the ideograms highlights additional features such as specific introgressions of cAus origin on chromosomes 1 and 10 among subtropical GJ (*Geng*/Japonica) types in Bhutan (upper portion of the GJ-sbtrp zone) (fig. 1) and specific introgressions of XI (*Xian*/Indica) origin on chromosomes 1, 2, 4, and 10 among tropical forms from China, Japan, and Korea only (lower portion of the GJ-trp zone in fig. 1). Among Indica accessions, the XI-2 bear the largest proportion of pure discordant assignments (median 25.4 million bp), followed by XI-1B and XI-admx (median 14.3 and 11.6 million bp, respectively) and far above the other two major groups XI-1A and XI-3 (medians 6.9 and 6.2 million bp, respectively). This distribution is mostly dependent on cAus assignments (fig. 2). The proportion of segmental cAus genome in the XI-2 genome rises above 5%.

Concerning outlier classifications, the majority of those found in CRGs are finely dispersed with no particular structure, emulating the sporadic assignments observed during simulations. However, beyond the CRGs, outlier assignments appear concentrated in the cBasmati type, where they represent close to 13% of the genome on average (min = 9%, max = 25%), with large segments consistently assigned to the outlier class (large segments of contiguous black windows, fig. 3).

We observe intermediate classifications to be nonrandomly distributed along the genome (figs. 3 and 4). Some cases clearly indicate full distribution overlap, presenting the same topography as observed in the simulation of unimodal

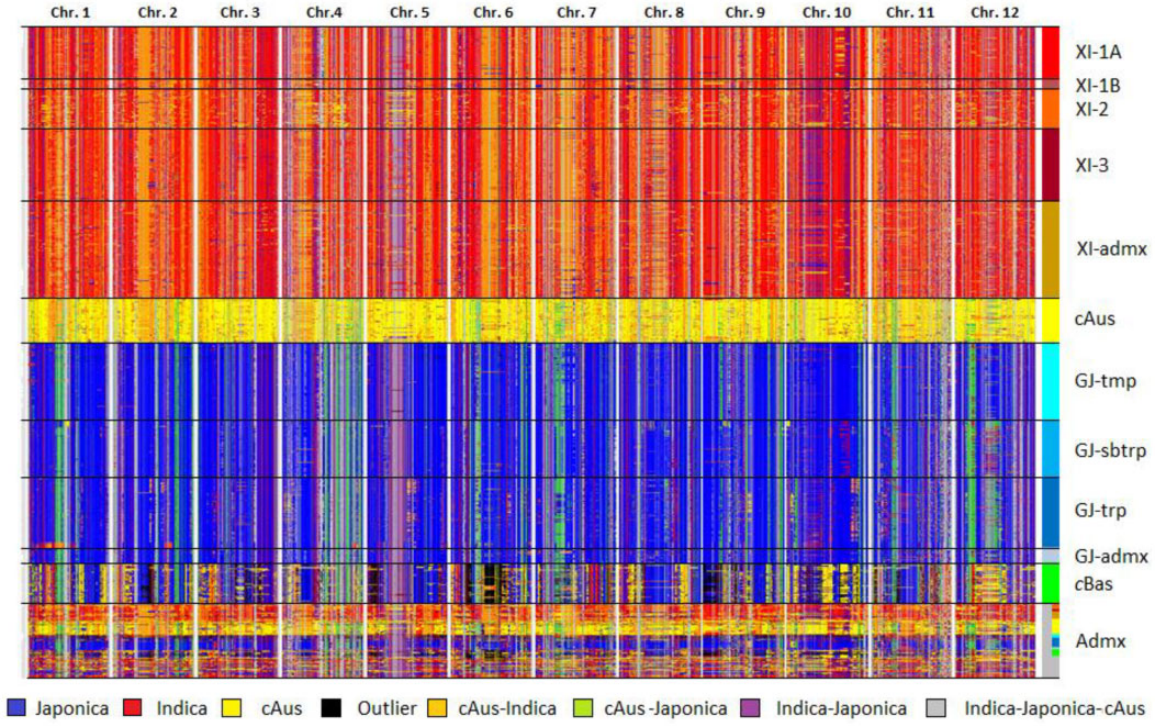

Fig. 1.—Complete genome ideogram of local classification across CORE Asian rice landraces. Patterns are organized per chromosome from left to right and the 948 accessions are arranged from top to bottom, organized first by reference groups and subgroups, then by geographic region of origin (not shown). Within the Admx, the accessions are arranged according to their classification in Wang et al. (2018).

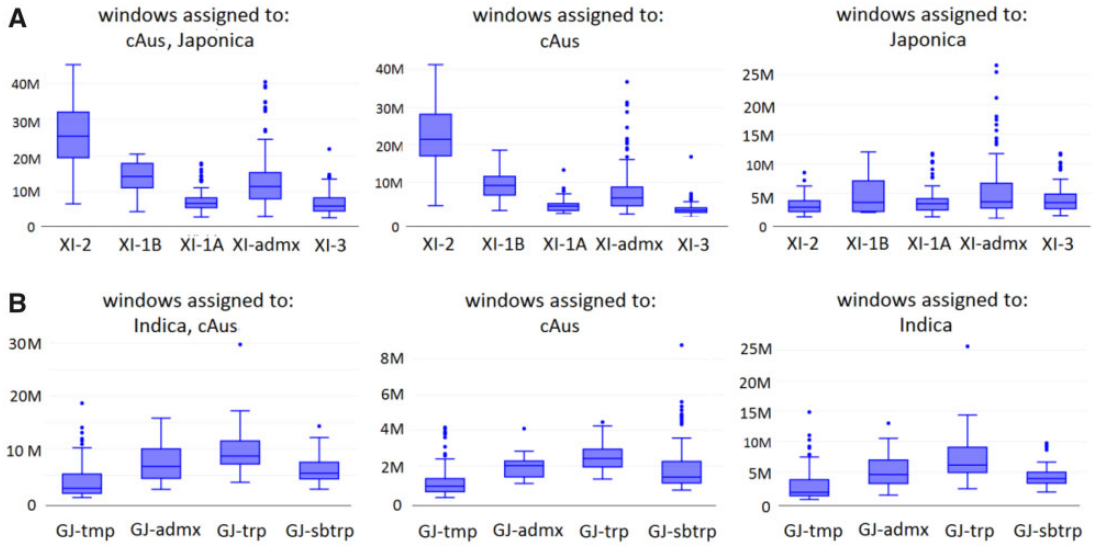

Fig. 2.—Genome coverage of pure assignments discordant with global classifications across subgroups of Indica (*A*) and Japonica (*B*). For each accession, the physical sizes of windows assigned to pure reference groups were summed across the genome. The distribution of total physical regions assigned to pure classifications discordant with the global classification by Wang et al. (2018) is analyzed across Indica subgroups (upper panel) and Japonica subgroups (lower panel). Sizes are given in millions of base pairs assigned (M).

distributions at low genetic distances (supplementary figs. S8 and S9, Supplementary Material online). Other cases are indicative of partial overlap, revealing the multimodal nature of CRG distributions (fig. 3).

The behavior of the KDE-based classification into outlier, intermediate and pure classes can be illustrated when its output is superimposed onto locally derived phylogenetic trees. A reclassification of haplotypes at some loci of interest was

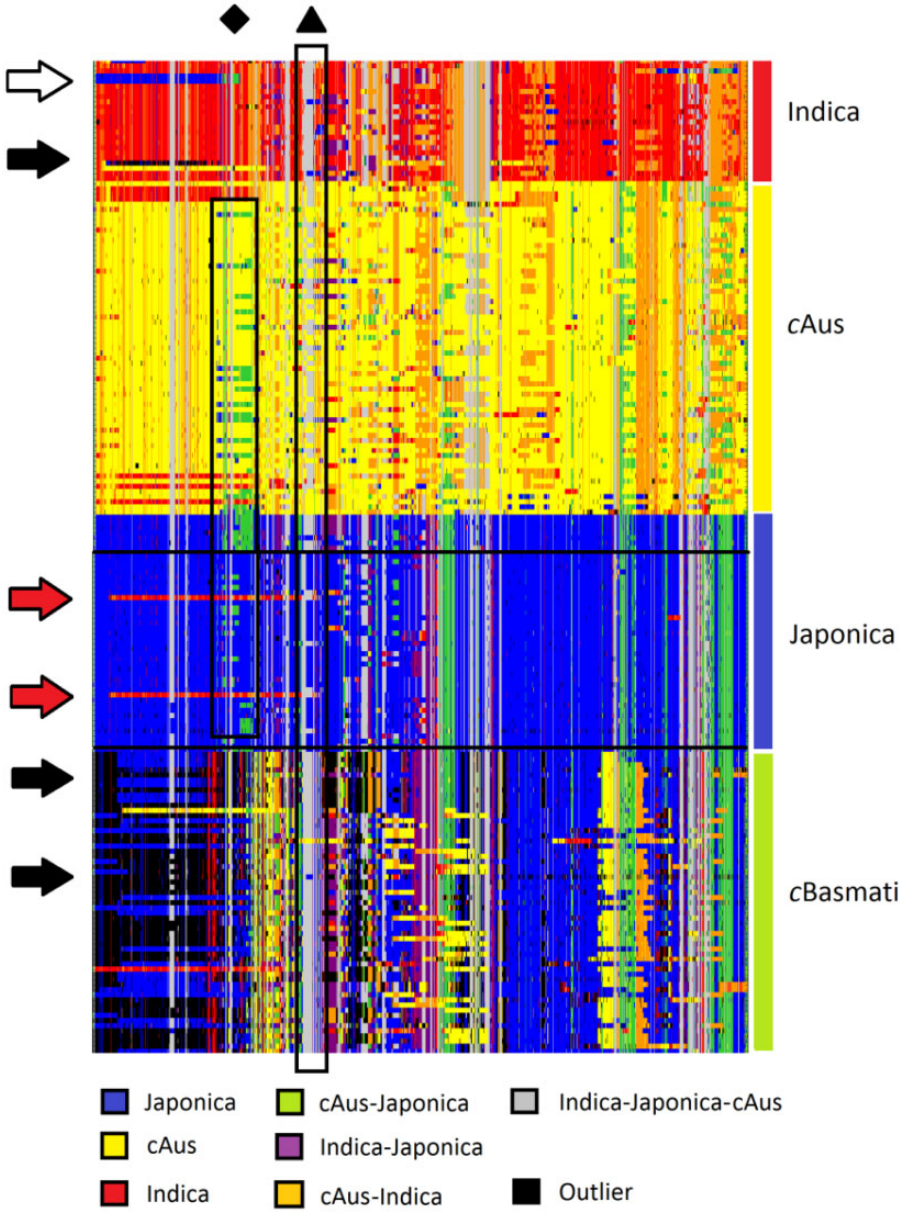

Fig. 3.—Extract of ideogram of local classification along chromosome 9 of Asian landrace rice accessions. White-filled arrow: example of local assignment contradictory to accession-specific global assignment; red-filled arrows: extended regions of shared assignment to Indica among two Japonica accessions MUANG TAY (IRGC 98382, GS 136100, Laos PDR) and MAK BOUAP (IRGC 30106, GS 132274, Laos PDR); black-filled arrows: examples of regions of consistent assignment to the outlier class—possible signature of the introduction of cryptic material (accession ARC 18061 [IRGC 47650, GS 127034, India] and cBasmati accessions at the bottom); black lozenge: example where the distribution of intermediate classification (Japonica–cAus in green) reveals the multimodal distribution of both CRGs; black triangle: example of a region of extended assignment to three-way intermediate class.

performed and plotted onto the corresponding neighbor joining graphs (fig. 5, supplementary figs. S15, S17, S19, and S21, Supplementary Material online). The output of the genome-wide analysis at those regions is displayed in figure 6*C* and supplementary figs. S16, S18, S20, and S22, Supplementary Material online.

Concerning the *bh4* locus (LOC_Os04g38660), we find that a major group of closely related haplotypes are consistently classified to the three-way intermediate group (fig. 5). However, local variation within that branch specific to a subgroup of Indica varieties results in their classification to their respective CRG. Pure cAus classification results from a

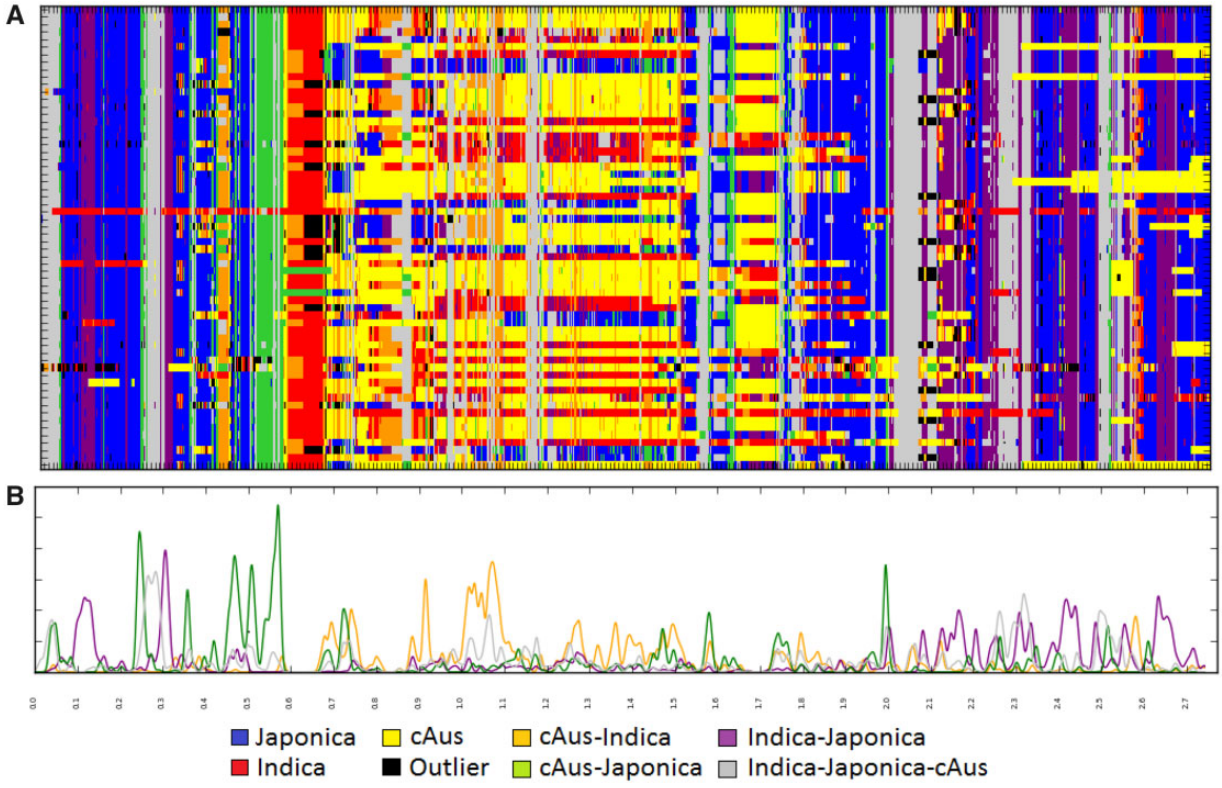

FIG. 4.—Summary analysis of local classification patterns of cBasmati genomes along chromosome 12. Genome-wide classification of local haplotypes into association-informative classes provides a platform for improved data analysis. Patterns of classification are explored for biological significance. (*A*) Ideogram representation of classification of cBasmati genomes (Wang et al. 2018) across chromosome 12. (*B*) Density of intermediate classifications across cBasmati accessions for Chr12.

distance to other groups that is not appreciable through ideogram analysis. Two distinct sources of outlier classification are visible: A small branch of peripheral Japonica haplotypes that are classified as outliers; a small group of cAus and cBasmati haplotypes isolated from the main bodies of variation.

The *qSH1* locus (LOC_Os01g62920) reveals a strong structure where the global classifications are well represented (fig. 5). Two large clusters corresponding to Indica and Japonica are clearly visible. Two other branches can be seen associating Indica and cAus, either separated (two subbranches of cAus and one of Indica), leading to their classification as pure CRGs, or intermingled, leading to an intermediate local assignment (orange group). Local Indica assignment puts together groups of haplotypes that are likely to have a polyphyletic origin. Finally, isolated and peripheral haplotypes can be observed to fall into the outlier class. Among these we identify the cBasmati, which connect at the origin of the main branch of Japonica variation at this locus.

Reference Overlap across Rice Genes

A median overlap score was estimated for each gene in the MSU7 database. Because of the way individual overlap scores were calculated, by summing pairwise comparisons of reference *P* values, values of 0 represent minimal ambiguity between the three references, values closer to one indicate the similarity between one pair of reference *P* values, and values close to three that similar *P* values were obtained from three reference distributions. The median was taken across CRG accessions as a measure of the overall degree of overlap at a given window, and averaged across windows when more than one window was considered.

Mean Shift clustering (Comaniciu and Meer 2002) of median gene overlap scores across genes revealed three non-overlapping groups with means 0.04 (55.4% of genes), 0.87 (32.7%), and 2.62 (12%), respectively (fig. 6*A* and *B*). We searched these groups for genes whose structure has been analyzed previously in the literature (fig. 6*C*). In the first group (mean = 0.04, no distribution overlap), we find the genes *qSH1* (LOC_Os01g62920, Zhang et al. 2009), a map-based cloned gene involved in seed shattering, *CHL1* (LOC_Os03g59640, Zhang et al. 2006) and *CHL9* (LOC_Os03g36540, Zhang et al. 2006), both involved in chloroplast development and photosynthesis. The first has been described as possessing a Japonica allele preventing the formation of the abscission layer and a consequent reduction in seed shattering, the latter two can be considered house-keeping genes in a photosynthetic organism. All three

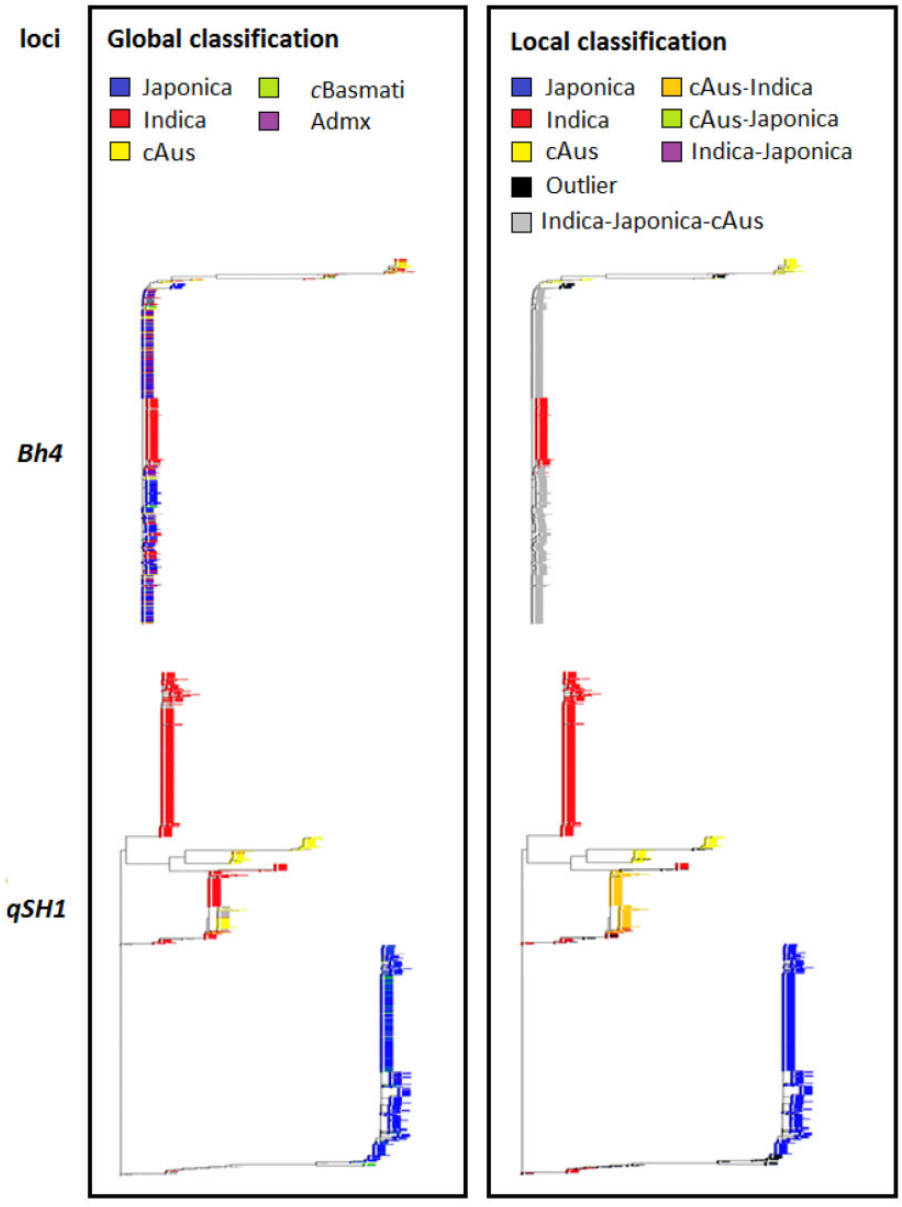

FIG. 5.—Core rice variation at *Bh4* (LOC_Os04g38660) and *qSH1* loci (LOC_Os01g62920) under global classification (left) and local, KDE-based classification (right). Regions encompass 5-kb upstream and downstream of the gene. Trees were constructed through Neighbor-Joining using the software DARWIN and a simple genetic dissimilarity index. (*A*) *Bh4* locus, 195 polymorphic SNPs identified between positions 22964845 and 22975964 of chromosome 4. (*B*) *qSH1* locus, with 165 polymorphic SNPs identified between positions 36440019 and 36454951 of chromosome 1.

display allelic variation matching the global classification scheme (fig. 6C-I). In the second group (mean = 0.87, pairwise distribution overlap), we find genes *rc* (LOC_Os07g11020, Sweeney et al. 2007), responsible for the red pericarp of some accessions, and *Waxy* (LOC_Os06g04200, Khush et al. 1984, see Sano et al. 1986), whose variation is responsible for the glutinous grain of some varieties. Both these genes determine traits that are easily assessed by humans and are subject to local preferences. Their classification into the second group is indicative of partial loss of structure. The surrounding regions display a predominance of two-way intermediate classifications (fig. 6C-II). In the third group (mean = 2.62) are found genes *Bh4* (LOC_Os04g38660, Zhu et al. 2011), responsible for the black hull, *OsC1* (LOC_Os06g10350, Saitoh et al. 2004), involved in apiculus coloration of the rice seed, and *sh4* (LOC_Os04g57530, Zhang et al. 2009), involved in grain shattering. The former two bear color-associated alleles predominant in wild rice and found only seldom in cultivated rice, generally straw-white seeded, whereas the latter is

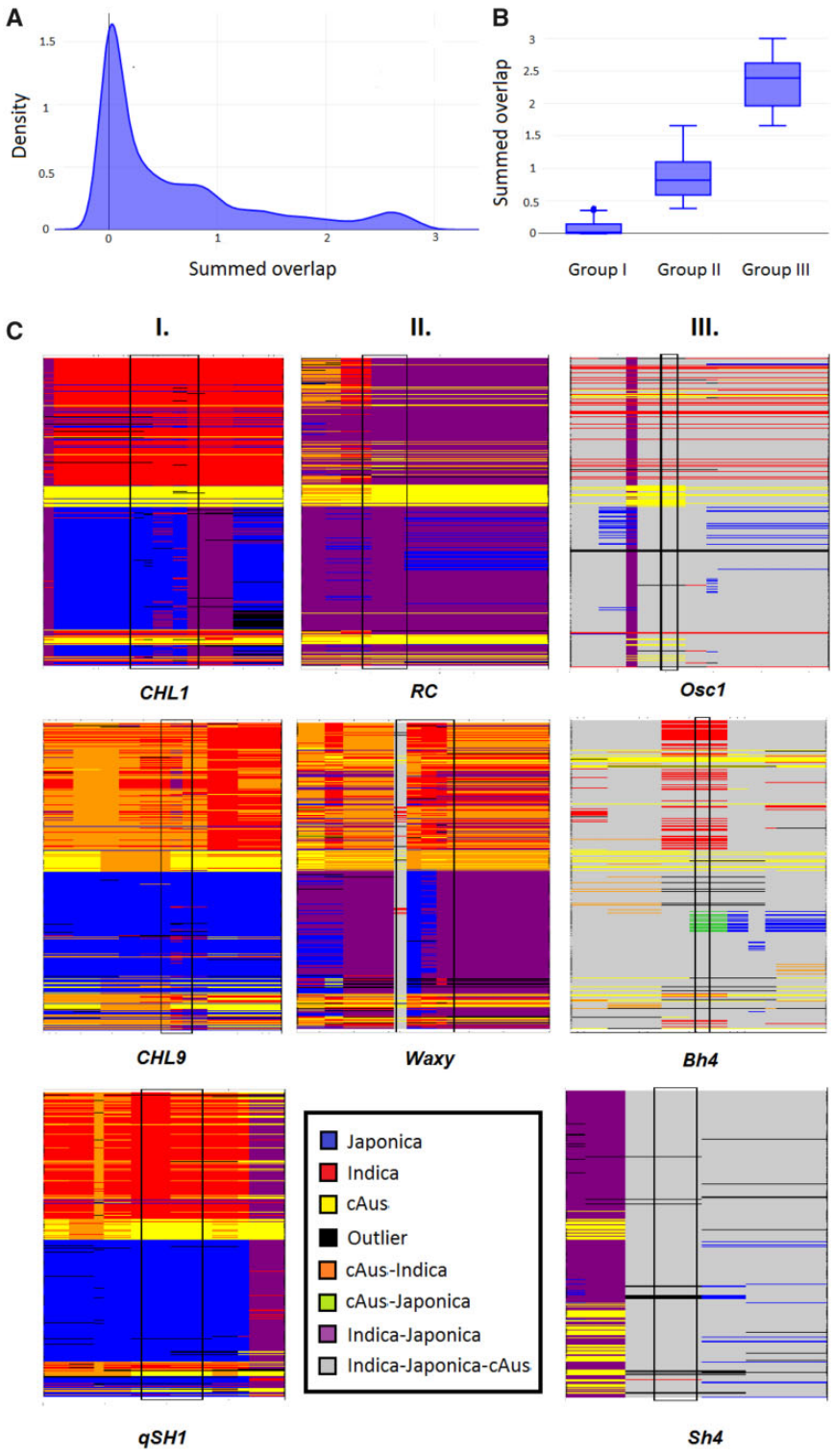

Fig. 6.—Summed *P* value overlap and local genomic classification of genes in the MSU7 Rice Genome Annotation Data base. Overlap was measured by summing the minimum to maximum proportion across pairwise combinations of CRG-specific *P* values for each individual. The median of this overlap was taken across reference accessions. Each gene was indexed to every window overlapping with its MSU7 coordinates. Overlap values for gene specific windows were averaged. (*A*) Distribution of median overlap across genes. (*B*) Boxplot of gene overlap values across groups identified through Mean Shift clustering on all genes. (*C*) Ideograms of genes of interest selected from within each of the groups identified through Mean Shift. MSU7 gene coordinates delimited by empty black rectangles.

responsible for a trait, seed shattering, which is central in domestication. Their classification into the third group is indicative of near-complete loss of structure and is accompanied by a predominance of three-way classifications in the surrounding regions (fig. 6C-III).

Discussion

We propose a representation of genomic diversity based on the distribution of patterns of similarity among reference entities. The KDE-based assignment into pure, intermediate and outlier classes is shown to provide a reliable summarization of the position of local haplotypes relative to the distributions of known groups, with intermediate classes functioning as indicators of distribution overlap. The final output is made more informative by these indicators. Nonetheless, more than one structure can lead to similar patterns of classification, a fact to be taken into account when analyzing the distribution of assignments across accessions.

Assignment Distributions

The simulation test of outlier identification showed that branches of variation even marginally represented by reference groups (roughly 3%) will be assigned to that population independently of genetic distance. In other words, assignment to a given class is not the same as assignment to a given evolutionary branch, but to the majority vote at any given cluster. The classes regarded as pure (corresponding to the three RGs) are not controlled so that they derive from any particular and single foundation event—this can be clearly appreciated in the overlay of classification over genetic distance at the *qSH1* locus for example. This makes the choice of reference accessions a key step in the analysis and is the reason why a relatively high threshold was set for the extraction of reference populations from the output of the global structure analysis (0.8, sNMF).

If these references are homogeneous and evenly differentiated along the genome, the patterns will exhibit relatively infrequent exceptions to the rule, illustrating limited and circumscribed exchanges. However, these references may be "imperfect." In the case of intrinsic heterogeneity (such as polyphyly), this internal structure will not be evident from the classification into pure classes. Concomitantly, if the references used fail to include major branches of diversity in the species (at an isolation threshold of 0.02 F_{st} from the nearest branch), this will result in assignment into the outlier class of nonreference accessions from those branches that may be included in the analysis.

However, if they are not homogeneous due to significant uneven exchanges along the genome, then this unevenness is promising matter for the search of differentially patterned genomic regions possibly bearing on evolutionary adaptation. Here, the nature of intermediate assignments must be considered. Classification into intermediate classes is not the product of specific evolutionary scenarios. The patterns retrieved by the current analysis alone do not conclusively identify genetic exchanges between reference groups, or the origin of the shared material. For example, we observe that local classification is reflected in our characterization of genes according to local assignment overlap and in line with our expectations regarding selection at those loci during domestication. At the same time, large sections of intermediate classification surrounding centromeres are also evident, suggesting hypervariability as the cause. Nonetheless, the discrimination between these and other hypotheses regarding reduced local structure should benefit from the accurate delimitation of these regions.

We believe this limitation to the interpretation of intermediate classifications is outweighed by the decrease in error rate at lower F_{st} values. For both our approach and that of *Loter* the shoulder of this distribution lay around an F_{st} of 0.03. Although we did not test the WINPOP method on simulated data, the relationship observed in our tests between accuracy of the KDE-based classifier and local pairwise genetic distance closely resembles that reported for that approach (Paşaniuc et al. 2009). It should be noted here that these three results reflect the "phase change phenomenon" observed by Patterson et al. (2006)—the differentiation threshold below which genetic structure is not noticeable. In this context, it is notable that out of the 22 methods of local ancestry inference reviewed by Geza et al. (2018), only four (WINPOP and LAMP, Paşaniuc et al. 2009; PCAdmix, Brisbin et al. 2012; SUPPORTMIX, Omberg et al. 2012) report estimates of accuracy versus local genetic distance between ancestral populations (although HAPAA is tested under a comprehensive gradient of generational drift between ancestral sources, Sundquist et al. 2008). In some cases, the global genetic distance between ancestral populations is discussed, but their range is limited to that of Human populations in the HGDP and HapMap data sets. In some other cases, genetic distance is calculated between fragments and sources of assignment only. All four cases where the relationship is described present similar results. The fact that of these four two are LD-based methods and two are not seems to indicate the limitation of this complement in solving this problem. In the case of rice, the importance of accounting for local genetic structure is strikingly clear: Under our analysis, an average of 46% of genomes of CRG accessions are assigned to one intermediate class or another (for 20% none of the reference groups is distinguishable).

In summary, the classification output provided here represents a reduction of the genetic diversity in this data set into an informed overview of local genomic proximity. In the future, we hope the partitioning of this variation will be used to parse local haplotypes for more detailed analyses.

Intermediate Classification

Analysis of global ideogram plots reveals regions of intermediate assignment to be localized in the genome. The pattern of classification into these classes among reference accessions in some cases points to multimodal distributions where at least one mode is shared by two or more reference groups. Although we cannot presently conclude on the origin of these shared structures, the higher estimated proportion of intermediate cAus–Indica versus Japonica–cAus or Japonica–Indica regions across the genomes of their respective accessions supports previous observations regarding their relative proximity (see Civáň et al. 2015). The discussion has been on whether these two modern populations were first sampled from independent wild populations or the same one, differentiating later. More recently, a study including 480 *O. rufipogon* accessions widely sampled in South and East Asia has shown their global structure to closely mirror that of modern cultivated varieties (Wang et al. 2017). Given the pervasiveness of spontaneous gene flow from cultivated varieties to their wild relatives, it is difficult to establish the cause of this resemblance. Although our study confirms the more recent common ancestry of cAus and Indica relative to Japonica (now controlled for recent exchanges between these groups), a more detailed, joint analysis of wild and domesticated rice diversity is needed to understand the order in which these events took place.

On a more general note, genetic structure at the subgenomic level can be impacted by selective pressures and various degrees of exchanges of genetic material. In more complicated scenarios—such as the present case study, one must often accept modern populations as a patchwork of ancient hybridizations, and thus to present varying evolutionary histories along the genome. In other words, regions of reduced structure, and their impact on local admixture analysis, are to be expected in such scenarios. The observations of the extent and distribution of these regions along the genome of rice should serve as a cautionary tale for future studies of local admixture. Although we hope fine tuning the analysis could reduce the size of these portions without loss of certainty, their extent itself demands caution.

Inter-Subspecific Exchanges

Patterns indicative of inter-subspecific local exchanges of genetic material are pervasive. Although in some cases these appear isolated, specific to a given variety, similar patterns of introgression are common among varieties of the same groups. Those varieties mildly affected by such exchanges are classified in the three principal groups, whereas those that are more strongly affected fall into the "admixed" remainder. They encompass the cBasmati varieties as well as other diverse varieties. Altogether this class is concentrated in the Indian Subcontinent, where all the cBasmati varieties are localized and where the frequency of other admixed varieties is about 5-fold higher than in the rest of Asia. The distribution of introgression patterns could indicate selection for introgressed phenotypes. Correlation with substructure could indicate a significant contribution to the process of divergence and isolation, if not explain it.

Genetic exchanges among the major centers of rice diversity have been the subject of multiple studies to date (Sun et al. 2002; Ishikawa et al. 2006; Sweeney and McCouch 2007). Even before a detailed quantification of their extent was attempted these exchanges were expected given the global nature of rice trade, the permeability of subspecies genetic barriers, and the obvious benefit of the transfer of population specific phenotypes of interest that has long spurred large-scale breeding programs such as undertaken by IRRI (Mackill and Khush 2018). Although the results of the more successful pedigrees in rice breeding programs are well known and have been propagated worldwide (Green Revolution, IR 8, IR 36, IR 64), the nature and frequency of contacts between modern genepools have never been fully quantified (Khush 1997).

It is to be expected that successful combinations of existing phenotypes would be selected. Those for which the extent of introgressions was minimized by generations of back crossing are found among our reference genepools. The remainder comprise the modern pool of heavily admixed varieties. We can see these to be a continuum of contributions from the three major groups of rice, as well as some differentiated material. Some coherence may exist among subsets of these rice hybrids, inviting deeper analysis and characterization of these varieties as potential key to understand rice reproductive barriers and the application of these data to successful breeding programs.

The Circum-Basmati

As subspecific hybridizations accumulate, it should come as no surprise that some combinations, hitting on phenotypes of wider economic interest, should spread and gain a more significant presence among modern cultivars. Included in the 3K Rice Genome data set were accessions pertaining to just such a seemingly hybrid but cohesive group of rice diversity, the cBasmati. Current knowledge places the cBasmati next to the Japonica, Indica, and cAus, as the fourth major group of genetically concordant cultivars (Glaszmann 1987; Garris et al. 2005). Their intermediate position relative to the Japonica and cAus, followed by the identification within this genepool of alleles private to both these populations had already led to the initial proposal and subsequent confirmation of their status as a hybrid group. Although supported by a smaller body of evidence, it had also been proposed there to have been contributions of Indica specific material to at least some individuals of this population (Jain et al. 2004). Our analysis supports a larger contribution of Japonica material to this genepool (average = 27.9%) and further highlights the contribution

of cAus (11.7%) and Indica (6%), not taking account of regions classified as intermediate.

In addition to the purported contributions from the three major genetic groups of rice, initial analysis of the 3K Rice Genomes data set by PCA showed the cBasmati to explain the majority of variation along a fourth axis while representing only around 2% of the total number of accessions (Wang et al. 2018). Our analysis outlines large regions of conserved assignment to the outlier class among accessions pertaining to this group. Although the exact nature of this material awaits a more detailed analysis, this narrows the possible localization of introgressed cryptic or wild material within cBasmati genomes to these regions. Recent archaeological findings (Bates et al. 2017) highlight the presence of domesticated rice among Indus settlements in the North West of India, then coexisting with wild forms of *Oryza nivara* species, in a period that started as early as about 3000 BCE, long before the arrival of the introduced Japonica forms around 2000 BCE. This region is notable for traditionally growing no Indica varieties, but only cAus, cBasmati, and Japonica varieties (Glaszmann 1988), thus pointing at the possibility of specific domesticates in the western part of the Himalayan foothills. On the eastern fringe of its distribution, in Myanmar, this group is known to form a specific subgroup which displays specific alleles at microsatellite loci as well as for BADH2 (Myint et al. 2012), the gene for aroma (Bradbury et al. 2008). Considering also the minute groups of deepwater/floating rices from Bangladesh and Northeast India (Glaszmann 1987; Bin Rahman and Zhang 2013), the Himalayan foothills appear as a swarm rich in hybrid forms between local domesticates, exogenous cultivars and possibly a series of local wild forms.

The Dynamics of Diversity

We have confirmed the global structure described since the advent of molecular markers and the finer structure recently revealed in the first analysis of the data set we have been using (Wang et al. 2018). We have provided a global estimate of the overlap among the major elements of this structure and highlighted genomic regions of specific interest for their homogeneity among all or some of the groups. These can be further analyzed in terms of gene content and selection signals in search for adaptation factors. Also in line with the analyses carried out earlier on these data, the patterns observed around genes known to bear variation connected to the process of domestication reflect phenomena of allele dissemination across varietal groups but also show the existence of group-specific alleles expected in diffuse domestication circumstances. We have also highlighted specific large genome segments in specific germplasm compartments that are obvious traces of recent introgression. Finally, with the example of the cBasmati varieties, we observe that variation that departs from the simplest model with a few foundation pillars can be very clearly exposed and taken as threads to unravel additional dimensions of the rice crop diversity.

On a more general line, should the process of genetic basis augmentation by occasional introgression be common rather than an exception, the pattern of diversity of the rice crop is likely to host many other genome segments of alien origin incorporated to the derivatives of initial domestication foundation pillars. Such segments would appear as outliers if the core reference sample were modified, thus calling for additional rounds of analysis at finer diversity scales beyond the work reported here. This novel vision justifies our initial methodological choice to focus first on patterns of structure before designing evolutionary model-based scenarios aimed at phylogenetic interpretation.

Conclusion

The history of a major crop such as rice, with its ancient origin and the evidence of profuse germplasm movements across the Asian continent, is undoubtedly complex and features numerous episodes of genetic exchange between diverse branches of a reticulate expansion. Retrospectively, current diversity patterns are first the result of recent admixture events among differentiated forms likely to still exist as such. Our approach provides a shallow analysis of the past, depicting traces of recent admixture and circumscribed introgression. This departs from a phylogenetic description, because this classification is not based on the absolute level of divergence but on the respective specificity—the relative prevalence of specific variation among distinct CRGs. In other words, we have developed a distribution-related assignment that is not immediately indicative of an actual phylogenetic origin. Although a complete depiction of the evolution of the species remains elusive, this first step gives access to a range of recent punctual exchanges with potential adaptive significance and provides possible keys for determining more finely which germplasm compartments, subgroups and geographic origins have been involved in these exchanges. Our approach also highlights the existence of outlier haplotypes to those that are found in the CRGs, and allows us to delimit those that can tentatively be considered to be of alien origin. Their distribution within the germplasm and along the genome can help determine new groups that could serve as additional references or localized contributions from exotic germplasm to be searched for among wild relatives. The example of the cBasmati varieties is illustrative of this case. These varieties exhibit genome portions that can be related to all the CRGs, yet about 13% of their genome appears to be composed of completely unique haplotypes. This explains why they appear as a distinct group in some studies and as an admixed population in others. This first step also enables us to "clean" diverse genomes from the direct consequences of recent exchanges. Pursuing this deconvolution will reveal deeper patterns that can then be analyzed, as indicative of

lineages that may have been lost or are underrepresented in current modern germplasm surveys.

Data Availability

All SNP genotyping data were obtained through the Rice SNP-Seek Database at http://snp-seek.irri.org/ (Mansueto et al. 2017). Raw *P* value estimates are stored in the figshare repository https://doi.org/10.6084/m9.figshare.7345991.v3. Complete ideogram representations of local classifications used for summary analyses are available in the figshare repository https://doi.org/10.6084/m9.figshare.7347053.v2. All scripts and summary analyses are accessible at the GitHub repository https://doi.org/10.5281/zenodo.2587930.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We thank Guillaume Martin and Aurélien Cottin for valuable advice and technical insight. We would also like to thank Nourollah Ahmadi for providing the first user test. This study was funded by CIRAD with support of the project Genome Harvest, reference ID 1504-006, through the "Investissements d'avenir" program (Labex Agro: ANR-10-LABX-0001-01), the project AdaptGrass, reference ID 170544IA, through the "Investissements d'Avenir" program (I-SITE MUSE: ANR-16-IDEX-0006) and the CGIAR Research Program on Rice Agrifood Systems (RICE). This work was supported by the CIRAD - UMR AGAP HPC Data Center of the South Green Bioinformatics platform (http://www.south-green.fr/).

Author Contributions

J.C.G., C.B., and J.D.S. conceived and designed the project. G.D. performed bioinformatics data management, J.D.S. developed analysis pipeline and simulations. J.D.S. and J.C.G. interpreted data. J.D.S., J.C.G., K.L.M., D.C., C.B., and J.B. wrote the manuscript.

Literature Cited

Aggarwal CC. 2016. Outlier analysis. 2nd ed. Yorktown Heights (NY): Springer Science & Business Media.

Bates J, Petrie CA, Singh RN. 2017. Approaching rice domestication in South Asia: new evidence from Indus settlements in northern India. J Archaeol Sci. 78:193–201.

Bin Rahman A, Zhang J. 2013. Rayada specialty: the forgotten resource of elite features of rice. Rice (N Y). 6(1):41.

Bradbury LMT, Gillies SA, Brushett DJ, Waters DLE, Henry RJ. 2008. Inactivation of an aminoaldehyde dehydrogenase is responsible for fragrance in rice. Plant Mol Biol. 68(4-5):439–449.

Brisbin A, et al. 2012. PCAdmix: principal components-based assignment of ancestry along each chromosome in individuals with admixed ancestry from two or more populations. Hum Biol. 84(4):343–364.

Civáň P, Craig H, Cox CJ, Brown TA. 2015. Three geographically separate domestications of Asian rice. Nat Plants 1:15164.

Comaniciu D, Meer P. 2002. Mean shift: a robust approach toward feature space analysis. IEEE Trans Pattern Anal Mach Intell. 24(5):603–619.

Diao L, Chen KC. 2012. Local ancestry corrects for population structure in *Saccharomyces cerevisiae* genome-wide association studies. Genetics 192(4):1503–1511.

Dias-Alves T, Mairal J, Blum M. 2018. Loter: a software package to infer local ancestry for a wide range of species. Mol Biol Evol. 35(9):2318–2326.

Frichot E, Mathieu F, Trouillon T, Bouchard G, François O. 2014. Fast and efficient estimation of individual ancestry coefficients. Genetics 196(4):973–983.

Garris AJ, Tai TH, Coburn J, Kresovich S, McCouch S. 2005. Genetic structure and diversity in *Oryza sativa* L. Genetics 169(3):1631–1638.

Geza E, et al. 2018. A comprehensive survey of models for dissecting local ancestry deconvolution in human genome. Brief Bioinform. 6 (29): bby044.

Glaszmann JC. 1987. Isozymes and classification of Asian rice varieties. Theor Appl Genet. 74(1):21–30.

Glaszmann JC. 1988. Geographic pattern of variation among Asian native rice cultivars (*Oryza sativa* L.) based on fifteen isozyme loci. Genome 30(5):782–792.

Glaszmann JC, Arraudeau M. 1986. Rice plant type variation: japonica-Javanica relationships. Rice Genet Newsl. 3:41–42.

Gravel S. 2012. Population genetics models of local ancestry. Genetics 191(2):607–619.

Henn BM, et al. 2012. Genomic ancestry of North Africans supports back-to-Africa migrations. PLoS Genet. 8(1):e1002397.

Ho P-T. 1956. Early-ripening rice in Chinese history by Ping-Ti Ho. Econ Hist Rev. 9:200–218.

Huang X, et al. 2012. A map of rice genome variation reveals the origin of cultivated rice. Nature 490(7421):497–501.

Ishikawa R, et al. 2006. Genetic erosion from modern varieties into traditional upland rice cultivars (*Oryza sativa* L.) in Northern Thailand. Genet Resour Crop Evol. 53(2):245–252.

Jain S, Jain RK, McCouch SR. 2004. Genetic analysis of Indian aromatic and quality rice (*Oryza sativa* L.) germplasm using panels of fluorescently-labeled microsatellite markers. Theor Appl Genet. 109(5):965–977.

Kato S. 1928. On the affinity of rice varieties as shown by fertility of hybrid plants. Sci Bull Faculty Agric Kyushu Univ. 3:132–147.

Kawahara Y, et al. 2013. Improvement of the *Oryza sativa* nipponbare reference genome using next generation sequence and optical map data. Rice 6:3–10.

Khush GS. 1997. Origin, dispersal, cultivation and variation of rice. Plant Mol Biol. 35(1-2):25–34.

Khush GS, Singh RJ, Sur SC, Librojo AL. 1984. Primary trisomics of rice: origin, morphology, cytology and use in linkage mapping. Genetics 107:141–163.

Mackill DJ, Khush GS. 2018. IR64: a high-quality and high-yielding mega variety. Rice (N Y). 11(1):18.

Mansueto L, et al. 2017. Rice SNP-seek database update: new SNPs, indels, and queries. Nucleic Acids Res. 45(D1):D1075–D1081.

Matsuo T. 1952. Genecological studies on cultivated rice. Bull Natl Inst Agric Sci Jpn. S3:1–111.

McVean G. 2009. A genealogical interpretation of principal components analysis. PLoS Genet. 5(10):e1000686.

Morinaga T. 1954. Classification of rice varieties on the basis of affinity. Jpn J Breed. 4: 1–14.

Myint K, et al. 2012. Specific patterns of genetic diversity among aromatic rice varieties in Myanmar. Rice 5(1):20.

Novembre J, Stephens M. 2008. Interpreting principal component analyses of spatial population genetic variation. Nat Genet. 40(5):646–649.
Oka HI. 1988. Origin of cultivated rice. Amsterdam: Elsevier.
Omberg L, et al. 2012. Inferring genome-wide patterns of admixture in Qataris using fifty-five ancestral populations. BMC Genet. 13:49.
Park DS, et al. 2015. Adapt-Mix: Learning local genetic correlation structure improves summary statistics-based analyses. Bioinformatics 31:i181–i189.
Paşaniuc B, Sankararaman S, Kimmel G, Halperin E. 2009. Inference of locus-specific ancestry in closely related populations. Bioinformatics 25:i213–i221.
Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. PLoS Genet. 2(12):e190.
Pedregosa F, et al. 2011. Scikit-learn: machine learning in Python. J Mach Learn Res. 12:2825–2830.
Perrier X, Jacquemout-Collet J. 2006. DARwin software. Available from: http://darwin.cirad.fr/, last accessed May 01, 2016.
Saitoh K, Onishi K, Mikami I, Thidar K, Sano Y. 2004. Allelic diversification at the C (OsC1) locus of wild and cultivated rice: nucleotide changes associated with phenotypes. Genetics 168(2):997–1007.
Sano Y, Katsumata M, Okuno K. 1986. Genetic studies of speciation in cultivated rice. 5. Inter- and intraspecific differentiation in the waxy gene expression of rice. Euphytica 35(1):1–9.
Scott DW. 2015. Multivariate density estimation: theory, practice, and visualization. 2nd edition, the first is from 1992. New Jersey:Wiley. https://books.google.fr/books? id=pIAZBwAAQBAJ.
Silverman BW. 1998. Density estimation for statistics and data analysis. 1st ed. New York: Routledge.
Slatkin M. 1991. Inbreeding coefficients and coalescence times. Genet Res. 58(2):167.
Sun C, Wang X, Yoshimura A, Doi K. 2002. Genetic differentiation for nuclear, mitochondrial and chloroplast genomes in common wild rice (*Oryza rufipogon* Griff.) and cultivated rice (*Oryza sativa* L.). Theor Appl Genet. 104(8):1335–1345.
Sundquist A, Fratkin E, Do CB, Batzoglou S. 2008. Effect of genetic divergence in identifying ancestral origin using HAPAA. Genome Res. 18(4):676–682.
Sweeney M, McCouch S. 2007. The complex history of the domestication of rice. Ann Bot. 100(5):951–957.
Sweeney MT, et al. 2007. Global dissemination of a single mutation conferring white pericarp in rice. PLoS Genet. 3(8):e133.
Wang H, Vieira FG, Crawford JE, Chu C, Nielsen R. 2017. Asian wild rice is a hybrid swarm with extensive gene flow and feralization from domesticated rice. Genome Res. 27(6):1029–1038.
Wang W, et al. 2018. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature 557(7703):43–49.
Wang YM, et al. 2005. Extensive de novo genomic variation in rice induced by introgression from wild rice (*Zizania latifolia* Griseb.). Genetics 170(4):1945–1956.
Zhang H, et al. 2006. Rice Chlorina-1 and Chlorina-9 encode ChlD and ChlI subunits of Mg-chelatase, a key enzyme for chlorophyll synthesis and chloroplast development. Plant Mol Biol. 62(3):325–337.
Zhang LB, et al. 2009. Selection on grain shattering genes and rates of rice domestication. New Phytol. 184(3):708–720.
Zhu B-F, et al. 2011. Genetic control of a transition from black to straw-white seed hull in rice domestication. Plant Physiol. 155(3):1301–1311.

Associate editor: Aoife McLysaght